Rethinking Quality Assurance for Crowdsourced Multi-ROI Image Segmentation

Xiaolu Lu[1], David Ratcliffe[1], Tsu-Ting Kao[1], Aristarkh Tikhonov[2], Lester Litchfield[2], Craig Rodger[1], Kaier Wang[2]

[1] Microsoft
[2] Volpara Health Technologies Ltd.
xiaolu.lu@microsoft.com, david.ratcliffe@microsoft.com, tsu.kao@microsoft.com, craig.rodger@microsoft.com,
lester.litchfield@volparahealth.com, aristarkh.tikhonov@volparahealth.com, kyle.wang@volparahealth.com

Abstract

Collecting high quality annotations to construct an evaluation dataset is essential for assessing the true performance of machine learning models. One popular way of performing data annotation is via crowdsourcing, where quality can be of concern. Despite much prior work addressing the annotation quality problem in crowdsourcing generally, little has been discussed in detail for *image segmentation* tasks. These tasks often require pixel-level annotation accuracy, and is relatively complex when compared to image classification or object detection with bounding-boxes. In this paper, we focus on image segmentation annotation via crowdsourcing, where images may not have been collected in a controlled way. In this setting, the task of annotating may be non-trivial, where annotators may experience difficultly in differentiating between regions-of-interest (ROIs) and background pixels. We implement an annotation process on a medical image annotation task and examine the effectiveness of several in-situ and manual quality assurance and quality control mechanisms. Our observations on this task are three-fold. Firstly, including an onboarding and a pilot phase improves quality assurance as annotators can familiarize themselves with the task, especially when the definition of ROIs is ambiguous. Secondly, we observe high variability of annotation times, leading us to believe it cannot be relied upon as a source of information for quality control. When performing agreement analysis, we also show that global-level inter-rater agreement is insufficient to provide useful information, especially when annotator skill levels vary. Thirdly, we recognize that reviewing all annotations can be time-consuming and often infeasible, and there currently exist no mechanisms to reduce the workload for reviewers. Therefore, we propose a method to create a priority list of images for review based on inter-rater agreement. Our experiments suggest that this method can be used to improve reviewer efficiency when compared to a baseline approach, especially if a fixed work budget is required.

1 Introduction

Recent advances in deep learning often depend on large-scale training datasets, and many applications require these to be high quality to obtain satisfying results (Freeman et al. 2021; Sambasivan et al. 2021).

One prominent way to create large labelled training datasets is to source them via crowdsourcing by engaging third party annotators on commercial platforms, such as Amazon Mechanical Turk[1]. In a crowdsourcing setup, participating annotators may have different backgrounds and varying skills levels, and a major concern of this approach is the resulting annotation quality. There are various studies on quality control and assurance in crowdsourcing tasks (Daniel et al. 2018). Broadly, these can be categorized into two types: (i) in-situ methods which include agreement analysis, such as computing Krippendorff's α (Krippendorff 2011) or telemetry analysis (e.g., annotation time analysis), and (ii) manual methods which require the involvement from experts and task owners, including the "shepherd mechanism" (Dow et al. 2012) or manual review (Ørting et al. 2020). Similarly, research in computer vision often leverages large-scale annotated datasets, such as ImageNet (Deng et al. 2009) or MS COCO (Lin et al. 2014). While much progress has been made in model development, researchers and practitioners are aware of the critical role that high quality annotated datasets play (Freeman et al. 2021), especially for challenging tasks such as medical image segmentation (Wang et al. 2022; Ji et al. 2021).

However, despite the adoption of recommended practice around general crowdsourcing setups for computer vision annotation tasks (Kovashka et al. 2016; Ørting et al. 2020), there are still gaps in quality control and assurance methods for *image segmentation* tasks. Image segmentation annotation differs from other computer vision annotation tasks such as classification, where labels are assigned to whole images or specified ROIs, or object detection where annotations are typically made with rectangular bounding-boxes. ROIs in segmentation tasks can be arbitrary polygonal regions. In this way, annotators must often perform a classification assertion on each pixel in an image to differentiate whether they belong to an ROI. Many of the quality control and assurance methods proposed in the crowdsourcing literature are not easily adaptable to image segmentation tasks, especially when images are collected from unconstrained environments (Lampert, Stumpf, and Gançarski 2016) such as medical images, or remote sensing. Detailed findings and experiments for crowdsourced segmentation annotation are not often discussed (Ørting et al. 2020).

This paper considers two types of quality control and as-

[1] Amazon Mechanical Turk: https://www.mturk.com/

surance mechanisms. The first of these types is in-situ redundancy analysis and annotator behavioral analysis. Redundancy analysis assesses the quality of multiple annotations made over the same instances, and is widely adopted in almost all types of crowdsourcing tasks. Specifically, *inter-rater agreement (IRA) analysis* is often used to assess annotator reliability across sets of labels, and can be used for quality control. Multiple labels which may be noisy can also be merged (Warfield, Zou, and Wells 2004) or directly incorporated into the modeling process (Jungo et al. 2018). In this paper, we consider the former scenario which is appliable for assessing *evaluation* datasets, whereas the latter methods are better suited to building *training* datasets. Specifically, we explore how to adopt Krippendorff's α (Krippendorff 2011), a common tool for measuring IRA, for quality assurance. We demonstrate that image-level Krippendorff's α values provide insufficient information for assessing label quality when multiple ROIs are present. Therefore, we propose computing localized Krippendorff's α values in addition to image level to gain a comprehensive understanding of annotation label inconsistency, which can then be provided back to annotators with a view to improving their output quality. Annotator behavioral analysis is another mechanism for quality control (Pei et al. 2021) which tries to correlate annotation quality with monitored telemetry such as annotation elapsed time, or number of clicks. In our work, we also explore whether annotation time is indicative of quality, especially for image segmentation, as it is rarely reported in existing work in the same setting.

The second type of quality control and assurance we consider requires external effort, relying on domain experts and task owners. While being more expensive than the aforementioned in-situ approaches, they are critical for many domain-specific annotation tasks such as medical data analysis (Ørting et al. 2020). For example, the shepherd mechanism (Dow et al. 2012) asks domain experts to interact with annotators to assist them in understanding definitions of ROIs clearly, and is critical when dealing with images collected in an unconstrained setting. However, most crowdsourcing workflows are optimized with respect to the experience of the annotators or task owners, and supervisory or reviewer-based roles are often overlooked.

In our work, we propose and test a simple yet effective way to help prioritize images for reviewers to assess, which our experimental results suggest can support the review process when the review budget is limited. Given the current gaps in the literature, we anticipate that our observations provide insights for future annotation work not only in medical image annotation, but other similar use cases, especially when non-expert annotators are involved.

2 Related Work

Quality Control and Assurance in Crowdsourcing. The quality of crowdsourced annotations is determined by a variety of factors, including ambiguities in annotator guidelines (Chang, Amershi, and Kamar 2017), or in the task itself which may be subjective, such as relevance judgements in information retrieval (Kutlu et al. 2018). Ensuring that crowdsourcing produces high quality labels is still an active research area. A comprehensive survey by Daniel et al. (2018) summarizes recent work in quality control and assurance for improving crowdsourced label quality, where it is observed that most all components and roles involved in crowdsourcing can have an impact on overall quality. Although there are many methods proposed for improving label quality, not all of them are generally applicable across domains. Understanding the extent to which some popular quality control methods provide useful insight in certain high stakes domains remains a largely unexplored area (Sambasivan et al. 2021), and we provide an overview in the next section. Arguably, most quality control techniques are implemented during the annotation process and roughly fall into one of three categories: (1) gold standard (e.g., (Kazai and Zitouni 2016)), which compares labels from annotators to ground truths; (2) redundancy analysis (e.g., (Waggoner and Chen 2014; Sheng, Provost, and Ipeirotis 2008)), which analyses the aggregated outputs from multiple annotators over common instances; and (3) annotator behavioral analysis (e.g., (Pei et al. 2021)), which tries to understand connections between annotator behaviors (e.g. annotation time) and annotation quality.

Besides starting a crowdsourcing annotation task with clear and well-written guidelines, "Train People" (Daniel et al. 2018) is another quality assurance mechanism which supports continuous learning by providing feedback to annotators throughout the entire process. Dow et al. concludes this strategy often yields better quality results. Involving domain experts or experienced annotators in the annotation process is another key strategy adopted in many high-stakes applications, such as the medical domain. Evaluating annotation quality for medical image segmentation is often done by incorporating reviewers in the annotation process. However, how such images are selected for review is not often detailed in the literature, but strategies include exhaustively reviewing all images (e.g., (Wang et al. 2017; Amgad et al. 2019)) or only a random sample (e.g., (Wang et al. 2022)). The workload of exhaustively reviewing all images can be time-consuming, as the image segmentation task itself is difficult, and one image may contain many annotations. There is currently limited work addressing this issue.

Inter-Rater Agreement in Image Segmentation Annotation Tasks. Inter-rater agreement (IRA), or annotation consistency, is commonly implemented in crowdsourcing tasks as an indicator of annotation quality. Many computer vision annotation tasks such as object-detection (Lin et al. 2014) or classification (Kovashka et al. 2016) also adopt this mechanism. However, few studies provide detailed descriptions on how inter-rater agreement is used for image segmentation tasks. One reason may be cost: due to the complexity of image segmentation annotation, many studies do not seek to acquire more than one annotation per image or instance (Lin et al. 2014; Lampert, Stumpf, and Gançarski 2016). In domains such as medical imaging or satellite imagery analysis where the annotation task can be complex and ambiguous, it is common to acquire multiple annotations for the same instances. Moreover, studies in these domains reveal high variability in inter-rater agreement analysis (Lam-

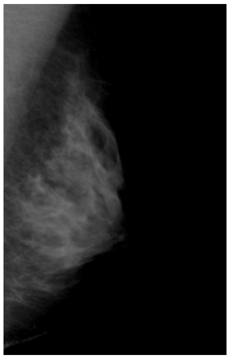


Figure 1: In the medical imaging domain, identifying anomalous features such as cancerous tissue or calcification in mammograms via segmentation can be challenging, often requiring domain expertise or deep familiarity with the task.

pert, Stumpf, and Gançarski 2016; Ribeiro, Avila, and Valle 2019; Krause et al. 2018) and subsequently focus on inferring ground truth from combining multiple labels. For example, STAPLE (Warfield, Zou, and Wells 2004) is a prominent annotation fusion algorithm for image segmentation tasks, which estimates the ground truth and annotator performance simultaneously using the Expectation-Maximization algorithm. However, such algorithms have revealed weak performance in some benchmarks (Gurari et al. 2015), and can lead to overly confident estimations of model performance when trained on fused annotations.

Recent development (Ji et al. 2021) try to leverage multiple labels directly in the modeling process and treats disagreement among raters as uncertainties. While noisy labels can successfully be used for training models with good performance, developing a gold standard test set with little ambiguity is nonetheless required for concluding the actual performance of a model.

Our work focuses on how to best measure the inter-rater agreement in the annotation process for image segmentation tasks with multiple annotators. We describe how the agreement value can be fed back to the crowdsourcing process in order to improve annotation quality. The closet published work to ours appears to be (Lampert, Stumpf, and Gançarski 2016) in which the authors examine the inter-rater agreement via the F1 score and evaluate how it impacts final model performance. Other work considered using Cohen's κ (Ribeiro, Avila, and Valle 2019) to measure agreement. However, when measuring the agreement for image segmentation, these methods consider the annotation task as one of binary classification task, with ROIs belonging to the positive class, and background pixels being negative. Our work demonstrates the ineffectiveness of this method in light of providing feedback to annotators, and proposes and analyses a modified version.

3 The Annotation Workflow

3.1 Background

Task and Data. We consider those segmentation annotation tasks whereby annotators are required to define and label arbitrarily complex polygonal ROIs in images, and where ROIs are not easily distinguishable by non-experts. This task requires annotators to both clearly understand the definition of ROIs and to use polygonal drawing tools to precisely capture which pixels belong to ROIs as opposed to background pixels. One such complex example is shown in Figure 1, where the task may be to highlight anomalous areas (e.g. cancerous, or calcified) within a mammogram X-ray image. In this paper, we refer to annotated ROIs as *masks*. Such masks may also be associated with a number of categories or labels (e.g., malignant, or benign), and although we focus on the use of a single category, we believe our work applies more generally to tasks with an arbitrary number of labels. Images such as mammograms as shown in Figure 1 can be considered an example of those collected in unconstrained environments with high variability, similar to the work of Lampert, Stumpf, and Gançarski, as they can be sourced from a number of different environments like hospitals or clinics, from various imaging sensors with varying settings and sensor angles, and taken of different patients with various physical characteristics. Unlike natural images of scenes where annotators may be required to identify and segment commonly well-understood ROIs such as cars or people, the ROIs in such medical images are often not easily distinguishable by non-experts, increasing the task difficulty.

Roles. We identify three roles involved in the overall annotation process: *task owners*, *annotators*, and *reviewers*. *Task owners* and *annotators* are no different from those in most other crowdsourcing tasks. *Task owners* are responsible for defining the crowdsourcing tasks, producing clear annotation guidelines, and the definition of ROIs for annotators to follow. *Annotators* are recruited to perform the annotation task, and we assume have some baseline level of understanding of the task and will not act maliciously (e.g., provide random annotations). *Reviewers* are often seen in difficult annotation tasks (Ørting et al. 2020), and can be experienced annotators or domain experts. For example, radiologists familiar with mammograms, or geographers who specialize in remote sensing. A major difference between reviewers and annotators besides expertise is their cost: hiring expert reviewers can be much more expensive than hiring annotators, however expert review is often critical to achieve high quality annotation datasets.

3.2 An Overview of the Annotation Workflow

Based on prior work in the crowdsourcing literature (Dow et al. 2012; Kovashka et al. 2016; Freeman et al. 2021), together with our task setup, we implement the annotation workflow shown in Figure 2 which consists of three main stages: Onboarding, Pilot, and Formal.

The Onboarding phase aims to familiarize the annotators with the task annotation itself, together with the platform and tools they are required to use. In this phase, annotators receive the task guidelines along with a small set of images, which they annotate. These images are pre-associated with gold-standard annotations and provide reviewers with the ability to assess annotator performance. Optionally, annotators can then be "pre-filtered" (Ørting et al. 2020; Lin et al. 2014) based on an acceptable performance threshold,

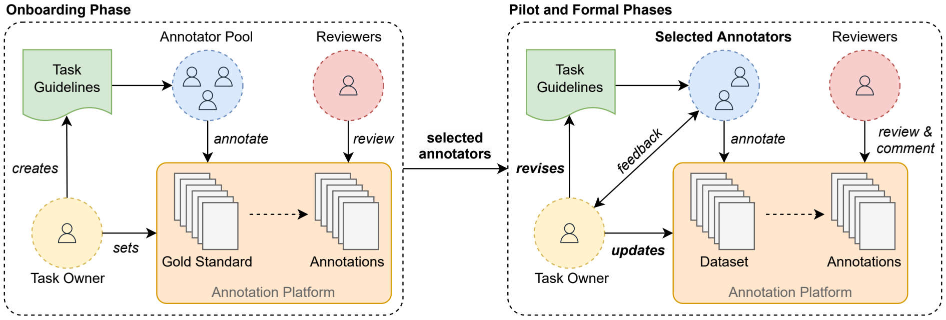


Figure 2: An overview of the annotation workflow considered in this work. Onboarding involves familiarizing annotators with the task over a small gold-standard dataset, then assessing their suitability for further participation. Selected annotators then participate in a Pilot phase where the task owner updates the task settings and revises the guidelines in preparation for the Formal annotation phase. In the Formal phase, the task owner updates the task settings and revises the guidelines as required.

and other factors such as task complexity, and annotation budget. Any annotators selected for further participation are then invited to participate in subsequent phases.

At this point in the workflow, a Pilot phase is often recommended (Daniel et al. 2018) before commencement of a formal annotation phase. The intention of a Pilot phase is to help the task owner assess whether the task setup and guidelines are correct, for example whether images are sufficiently prepared, and the description of ROIs is well conveyed. It also helps establish a communication channel between annotators, reviewers, and the task owner, which is the basis for implementing the "shepherd mechanism" (Dow et al. 2012) with feedback.

Once the Pilot phase in complete and the task settings and guidelines have been updated, the Formal annotation phase can begin. As shown in Figure 2, this phase is similar to the Onboarding and Pilot phases, in that selected annotators will work on the annotation task based on the updated task guidelines, and reviewers will assess the annotator outputs, deciding if annotations should be accepted or if they require revision. Instead of treating the task setup and guidelines as static, we incorporate the idea proposed by Freeman et al. (2021) that they should be updated throughout the annotation phases as required, including as the formal annotation phase progresses.

Lastly, it is possible that new annotators require onboarding midway through the execution of the annotation workflow, potentially after the Pilot phase has been completed. In this case, we expect that new Onboarding phases are conducted for any new group of annotators to prepare them to contribute, and can skip the Pilot phase if the task owners feel the workflow settings or task guidelines no longer need updating.

3.3 Quality Control and Assurance

External Quality Assurance. Figure 2 includes several external quality assurance methods, including reviewer/annotator interactions. We note that the review process provides for more than simply generating a binary decision (i.e. accept, or reject), as it also permits reviewers to convey informative or instructional comments to annotators, as motivated by Dow et al. (2012) as a "shepherding mechanism". By interacting with reviewers, annotators can better understand the task and the definition of ROIs. When applying this mechanism, each annotation phase (Onboarding, Pilot, and Formal) will have produced multiple versions of annotations for each image. Those annotations for which no further revisions are deemed required are treated as the "final" version and will be used in downstream tasks (e.g., training deep learning models). Our hypothesis is that, when annotators become sufficiently familiar with the task and our setup is adequately tuned, we would expect a significant number of annotations to be accepted without requiring further revision. By assessing annotator performance in review, task owners can identify any corner cases not adequately described in the existing task guidelines, and update them accordingly.

In-Situ Quality Assurance. As one of the widely adopted strategies for behavioral analysis, we choose to inspect annotation time spent per image. This is the most tangible signal, as almost all mainstream annotation platforms provide such statistics. Through analysis of annotation time, we aim to understand if there is any correlation between low quality annotations, the time spent on these images, and the annotation task difficulty. However, such behavioral analyses can only provide a weak signal to understand annotation quality. As per most other crowdsourcing methods, our main mechanism for in-situ quality assurance is redundancy analysis. Specifically, we rely on Krippendorff's α as it supports assessment over the results of multiple annotators, and handles missing labels (Krippendorff 2011). Computing a Krippendorff's α score for multiple segmentation annotations of the same image is not as straightforward as for other annotation tasks like classification. At first glance, it can be computed by treating each pixel in an image as a subtask. For example, in Figure 3, we consider the annotation and ground truth as two `0-1` sequences, where "`0`" means background pixel and "`1`" indicates some ROI. An overall image-level Krippendorff's α score can then be computed over these two sequences. However, this approach hides local differences which could be informative of annotator error. For example, in Figure 3, the image-level Krippendorff's α score is 0.73,

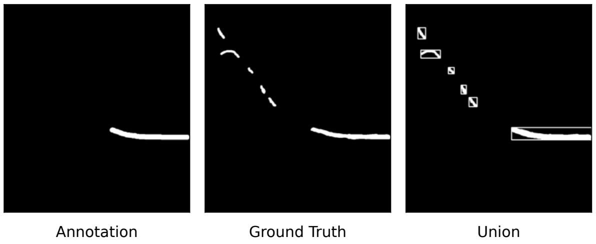


Figure 3: An example of an annotator's image mask and a ground-truth annotation marking highlighting ROIs, and their union with bounding-boxes around discrete connected regions. The underlying image has been omitted, and only ROI masks are shown. The image-level Krippendorff's α score between the annotator and ground truth masks is 0.73 (moderate agreement), while the average bounding-box-level Krippendorff's α score is -0.092 (disagreement).

which we might regard as "moderate agreement". However, clearly the annotator has missed many discrete regions when compared to the ground truth annotation. This error pattern suggests that the annotator may have had difficulty identifying certain ROIs in this image, instead of accidentally missing these areas. This is an important distinction, so in this paper, we propose to compute a localized bounding-box-level Krippendorff's α score in addition to the image-level Krippendorff's α score as a new signal for analysis. The bounding-box-level Krippendorff's α scores for each image are computed using the following steps:

1. Create a union of all annotation masks, and compute discrete connected component mask contours from these.
2. Generate a set of bounding boxes which localize all discrete contour masks. This can be seen in Figure 3, as the bounding boxes over discrete masks in the union.
3. For each bounding box, compute a Krippendorff's α score by treating multiple annotations within the bounding box region as separate `0-1` sequences. Consider the example shown in Figure 3, where we have six Krippendorff's α scores for the entire image.

The example shown in Figure 3 has an average bounding-box-level Krippendorff's α score of -0.092, indicating strong systematic disagreement between the annotator and ground-truth. This example highlights why a manual review step may be necessary for image segmentation annotation tasks involving multiple ROIs, because if image-level Krippendorff's α scores are used exclusively, we may be led to believe the annotation is of sufficient quality. However, when looking at the bounding box-level Krippendorff's α scores, we reach a different conclusion. There is no rule-of-thumb when determining quality based on image-level or bounding box-level Krippendorff's α scores. Instead, reviewers must interpret disagreements in localized regions where they occur. This can provide insight into new patterns or cases for ROIs which may not have been made clear to annotators through the task guidelines, which could be subsequently revised.

4 Prioritizing Images for Review

Thus far, we have discussed the necessity for manual review steps throughout the annotation workflow. However, if the volume of annotations is large, and if reviewer effort is limited due to high costs or a fixed budget, we are motivated to find those images for which review can be the most informative to annotators.

Random image selection for review is an easy and widely adopted strategy, however may cause reviewers to skip annotations with systematic problems, resulting in low quality annotations overall if not corrected with annotators. Instead, our approach is to tackle image selection for review as a *ranking* problem, which prioritizes those annotations we deem most likely to be informative when reviewed in top ranked positions. We interviewed our reviewers, and their preference is (paraphrased) as follows:

> "*At an early stage, image-level inconsistencies are more critical to inspect, as it is a step before determining an accurate ROI . . . The accurate localization of ROIs is also important, especially for building segmentation models . . .*"

Therefore, images with a low Krippendorff's α score, especially an image-level Krippendorff's α score, ought to be ranked higher in a priority list for review. We therefore propose three methods for review prioritization which can be classified into two main categories: *sorting-based* methods and *scoring-based* methods. Sorting-based methods only perform a simple sort operation over a number of "features", while scoring-based methods aggregate such "features" into a single numeric value as the basis for ranking.

Sorting-Based Methods. As reviewers prefer to assess images with low Krippendorff's α scores across annotations, a natural ranking method is to sort images by ascending Krippendorff's α scores. This guarantees that images will be sorted according to the level of disagreement computed at an image level.

For the same ROIs, we permit two possible labels, *certain* and *uncertain*, allowing annotators to mark areas for which they are unsure, primarily to focus expert review. For sorting by Krippendorff's α scores, we distinguish two cases, which we denote *strict* and *relaxed*. *Strict* scores ignore areas labelled as *uncertain*, and *relaxed* scores treat these as belonging to ROIs. We employ both *strict* and *relaxed* Krippendorff's α scores as primary and secondary sorting features, respectively. Besides these top two features, others are listed in Table 1 and can be used for tie-breaking. Namely, if a clear ranking is established for images using only a *strict* Krippendorff's α score, other features can be ignored. If using both *strict* and *relaxed* Krippendorff's α scores still results in ties, then we can perform tie-breaking by the level of uncertainties indicated by annotators (i.e., the unsure label% and total number of unsure pixels), as we assume that a high percentage of uncertain pixels indicates the annotator expeienced difficulty in segmentation, requiring manual review. The annotation area-related features are followed by those based on the intra-rater Krippendorff's α scores, which are the final rows of image-level features in Table 1, and are only considered if ties are still present after sorting based on all previous features. However, based on our observations in Section 3.3, we believe that image-level Krippendorff's α scores may not reflect the true severity of disagreement for this task. Therefore, we also propose a bounding-box-level sorting method, using features listed in Table 1. Unlike the image-level Krippendorff's α score which is a scalar value, the bounding-box-level score of an image may consist of multiple Krippendorff's α scores as each bounding box is associated with one Krippendorff's α score. Instead of using descriptive statistics such as mean or average to aggregate Krippendorff's α scores for each image, we break the Krippendorff's α scores into different ranges according to default thresholds (Krippendorff 2011) and then count the number of bounding boxes for which their Krippendorff's α scores fall into each range:

- Disagreement: Krippendorff's $\alpha \leq 0.1$;
- Low agreement: $0.1 <$ Krippendorff's $\alpha \leq 0.667$;
- Moderate agreement: $0.667 <$ Krippendorff's $\alpha \leq 0.8$;
- High agreement: Krippendorff's $\alpha > 0.8$.

The sorting operation then is performed similarly to the image level, whereby features are considered according to the row order listed in Table 1. Since our primary features are counting-based features, they have a higher chance of producing ties due to being integer values. Therefore, a tie-breaking method is required for bounding-box-level sorting. One way is to consider both the area of each bounding box and their Krippendorff's α scores. Let i be i-th bounding box, α_i its Krippendorff's α score, area$_i$ its area in pixels, and N be the total number of bounding boxes in the *union* of all annotations. We then define the weighted bounding-box area W_{bbox}, as:

$$W_{\text{bbox}} = \sum_{i=1}^{N} (1 - \alpha_i) \cdot \text{area}_i \tag{1}$$

As image size can vary, we further normalize W_{bbox} with the image size, and apply it as our last feature for bounding-box-level sorting.

Scoring-Based Methods. A limitation with sorting methods is that only one or two features will often be considered as primary and others as tie-breaking features. Therefore, we propose a ranking function that leverages bounding-box-level Krippendorff's α scores, as we believe it provides more detail when compared to image-level Krippendorff's α scores. Motivated by the literature in information retrieval (Singhal, Buckley, and Mitra 1996; Jones 1972), we first consider a pivoted normalization of W_{bbox}. For each image, we compute the total bounding-box area using $\sum_i^N \text{area}_i$, and for all images in the current batch, we then have an average bounding-box area which we refer to as avg_bbox_area. The pivoted normalization of W_{bbox} is computed as:

$$\text{pivoted-}W_{\text{bbox}} = \frac{W_{\text{bbox}}}{\sum_i^N \text{area}_i / \text{avg\_bbox\_area}} \tag{2}$$

Compared to normalizing with the image size, this approach down-weights images with large annotation areas compared to smaller ones. The intuition here is that when there is a large area of annotation that exceeds the average area, it is more likely to contain inaccurate labels. This may be less interesting to reviewers when compared to complete misses, or a large percentage of false positives or negatives, which potentially indicate missed ROI patterns or features.

In addition to the pivoted-W_{bbox}, we also consider an inverse document frequency (IDF)-like weighting scheme, which we refer to as W_N:

$$W_N = \ln\left(\frac{N}{\#\text{images with } N \text{ bounding boxes}}\right) \tag{3}$$

The final score of an image with multiple annotations is then computed using pivoted-$W_{\text{bbox}} \cdot W_N$. By sorting the score in descending order, we obtain a priority list for review.

5 Experiments

We conduct our experiments over a medical image segmentation task, as the images have been collected in unconstrained settings, and the ROIs are difficult to discern for non-expert annotators (as described in Section 3.1). Specifically, we explore three main questions in our experiments:

- How useful is external quality assurance for this task?
- How much information can in-situ quality assurance methods provide us in terms of label quality?
- How effective is the priority list we create for reviewers?

5.1 Experiment Setup

The Task. Breast arterial calcification (BAC) is the build-up of calcium deposits in the walls of arteries in breast tissue. These calcium deposits are often visible as high-contrast white areas in mammogram images due to their strong attenuation of X-rays. Recent findings have linked the presence of BAC with coronary artery and cardiovascular diseases (Iribarren et al. 2022; Lee et al. 2020), and the number of studies that localize and quantify BAC using image segmentation models is increasing. Identifying BAC deposits in mammograms to build training and evaluation datasets for segmentation models is a non-trivial task, and accurate localization of BAC requires training.

While patients with limited BAC deposits indicative of early-stage cardiovascular disease are likely to benefit most from diagnosis and treatment, BAC in these patients can be particularly difficult to identify. Over-diagnosis of BAC in patients is also undesirable, as it may lead to expensive and unnecessary referrals to cardiologists or medical intervention. Therefore, highly accurate identification and localization of BAC in patients is desirable. Given the nature and difficulty of the task, localization of BAC lends itself primarily to expert radiologists who are familiar with the condition. However, in order to train deep learning models for automatic segmentation of BAC deposits, the exclusive use of medical experts to construct large training datasets is often prohibitively expensive. This is why we are motivated to seek efficient ways of incorporating cheaper non-experts in the annotation process through crowdsourcing, together with medical experts in a way which reduces their overhead in annotation and review for the BAC segmentation task.

Image-Level		Bounding-Box-Level	
Feature	Sorting	Feature	Sorting
Strict Krippendorff's α	asc	# Disagreement boxes	desc
Relaxed Krippendorff's α	asc	# Low agreement boxes	desc
Unsure labels%	desc	# Moderate agreement boxes	desc
Total number of unsure pixels	desc	# High agreement boxes	asc
Annotation areas (relaxed) std.	desc	# Bounding boxes (N)	desc
Average annotation areas (relaxed)	desc	W_{bbox} (Eq 1) divided by image size	desc
Intra-rater Krippendorff's α strict	asc		
Intra-rater Krippendorff's α relaxed	asc		

Table 1: Features used in the sorting-based methods. "asc" refers to the feature being sorted in ascending order, and "desc" refers to the feature being sorted in descending order. The sorting is performed by considering features from top to bottom.

We establish a crowdsourcing task for annotating BAC deposits in mammogram images, and provide task guidelines including why the task is important, how to annotate, examples of good and bad BAC annotations, and easily confused non-BAC patterns.

Annotators can identify BAC deposits using polygonal drawing tools to produce masks over BAC areas, labelling them red if they are certain, or green if they are uncertain. Areas without BAC are to be left unmarked. They are also instructed that not all images may contain BAC, and that images may require manual adjustment of their contrast or brightness in order to improve BAC visibility. As the images are collected in an unconstrained way, it is not feasible to apply standardized image normalization beforehand.

Data. We use 1,195 mammogram images sourced from a breast imaging clinic, consisting of scans from 168 unique patients, and 298 individual studies. Each study consists of three or four scans of the same patient, sourced over time. All patients in these studies were considered to have BAC by an expert radiologist, however while most images contain BAC, some do not. To use this data in our research, we have obtained ethics approval, and all image data used in this paper is pre-processed and HIPAA[2]-compliant. Individual identifying patient information had been removed to protect patient privacy. All annotators and reviewers received these pre-processed images, ensuring they could not access any identifying patient metadata, and their judgement is based exclusively on image features.

Working directly on unprocessed mammogram images is difficult, especially for non-experts, as it requires manual adjustment of the image histogram or the application of complex transformations to make subtle BAC patterns more visible. In this paper, we use Volpara® software to generate so-called *pseudo density maps* for annotators to work on as opposed to the original images (Highnam et al. 2010). A pseudo density map is based on physical characteristics of breast tissue in an image, and exposes high density areas such as calcium deposits as bright white, with low density areas being dark, as shown in Figure 4. We randomly select 20 images from the 1,195 images and obtain gold standard annotations for these from expert annotators. This gold standard dataset is used for onboarding annotators, over which we can compute Sørensen-Dice coefficient (Dice/F1) scores to quickly ascertain image-level annotation quality.

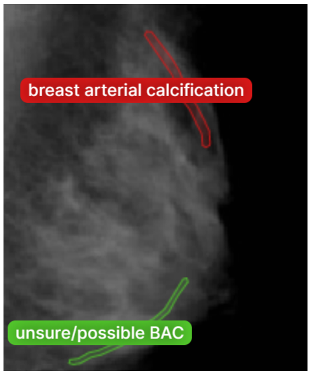


Figure 4: An example of a Volpara® pseudo density map with BAC annotations. Annotators identify BAC areas they are certain of in red, and low confidence areas in green.

Participants and Tools. In our study, we consider three groups of participants: A0, A1, and A2 (each with 5, 5, and 3 participants respectively), with each participant being familiar with the image data labelling process but having varying levels of expertise in interpreting mammograms and identifying and localizing BAC. Both A1 and A2 group annotators were familiar with interpreting mammograms, but only A1 annotators were expertly trained to identify BAC prior to the task. In contrast, A0 and A2 annotators were crowdworkers recruited from independent contracting firms paying at least minimum wage in each country of residence. As such, our experiments that explore the impact of different quality assurance approaches only consider these latter two groups.

In this paper, we use V7 Darwin[3] as the annotation platform. V7 Darwin provides several annotation tools like a brush, polygon, and auto-annotation tools for image segmentation. It also allows task owners to establish a workflow that involves an arbitrary number of review stages, which is useful for implementing a shepherd mechanism. Besides the

[2] HIPAA: The US Health Insurance Portability and Accountability Act of 1996.

[3] https://darwin.v7labs.com

annotation tool, we communicate with annotators through a Slack[4] team channel. When annotators had any questions or requests about their tasks, we were able to respond quickly.

5.2 Effectiveness of External QA Methods

Experiment Settings. In this experiment, we consider only externally recruited crowdsource annotators from A0 and A2 groups. Since our goal is to explore the effectiveness of the external quality assurance method of shepherding, we split the two groups of annotators to let them work on different annotation processes:

- A0 annotators only participated in Onboarding and Formal phases, receiving no interactions from reviewers. Reviewers were tasked only with assigning "reject" or "accept" to annotations after the entire annotation process was complete.
- A2 annotators participated in the full annotation process as described in Figure 2, which included an additional Pilot phase. During *all* annotation phases, annotators interacted with reviewers to better understand the reasons behind rejected annotations.

Note that the "acceptance" ratio reported in this paper is measured after the *first* round of iteration. There are 20 randomly sampled images in Onboarding phase, and 500 and 675 images in Pilot and Formal phases, respectively.

Initial Annotator Performance. Before exploring how annotators perform across different annotation phases, we use the Onboarding phase to assess their understanding of the task. A0 and A2 annotators initially participated in an Onboarding phase. Although both these groups of annotators claimed to have medical backgrounds and exposure to image segmentation tasks for machine learning, they were not initially familiar with identifying and localizing BAC areas in mammograms.

In their Onboarding phase, A0 and A2 participants were asked to annotate a small set of images.

Some annotators were not confident in carrying out the task, and were filtered out. The remaining annotators from groups A0 and A2 produced average Dice/F1 scores of 0.684 and 0.772 respectively, suggesting similar performance and assumed level of understanding of the task. However, as discussed in Section 3.3, we take caution in interpreting such image-level agreement scores, as these may not reveal fine-grained segmentation performance where multiple ROIs/BACs are present.

Shepherd Mechanism and Living Task Guidelines. In the Onboarding phase, we observe that annotators from both A0 and A2 groups have similar performance over the 20 images as per the acceptance ratio in Table 2. A0 annotators only have three more images accepted relative to the A2 annotators. However, in the Pilot phase, A2 annotators show a much improved acceptance ratio from 35% to 85%. We believe this is due to the frequent interactions between A2 annotators and reviewers, as although the acceptance ratio is low in the Onboarding phase, we believe that the annotators learn from feedback about reviewer-rejected annotations, helping them better understand how to identify ROIs/BACs. Further evidence is observed when comparing the final A0 and A2 acceptance ratio after the Formal phase, where A2 annotators achieve an acceptance ratio of 70% when compared to A0 annotators at 14%. For A0 annotators, the drop in acceptance ratio between Onboarding (50%) and Formal phases (14%) suggests errors were perpetuated across larger image batches.

Phase	Group	Images	Annotation Time (s)			Acc. Ratio(%)
			All	Accept	Reject	
Onboarding	A0	20	–	–	–	50
Onboarding	A2	20	–	–	–	35
Pilot†	A2	200	107±138	88±121	218±176	85
Formal†	A2	655	64±69	57±68	79±69	70
Formal†	A0	655	58±34	50±42	59±33	14

Table 2: Average annotation times (sec) with stdev for each image and acceptance ratio (Acc. Ratio%), denoting the proportion of accepted annotations overall. A † means annotation times between accepted and rejected is significantly different at $p = 0.05$ via a two-sided Welch t-test.

Reviewers also learn from this interactive process, as when they leave comments, they may observe patterns in mistakes made by annotators and can summarize them. We observed this amongst reviewers in the Formal phase where new error patterns were discovered which were not yet captured the guidelines, and explains why the acceptance ratio is lower in this phase. This leads to our design of a "living" task guideline document. Besides communicating directly with annotators on their mistakes, we add any common error patterns to the task guidelines. Annotators may then refer to these guidelines in subsequent work to avoid making the same mistakes. We also observed from the acceptance ratios that the interaction effort required by reviewers decreases as batches of images are annotated.

We conclude that both the shepherd mechanism and the living task guidelines can assist in quality assurance and in obtaining high quality annotations, both of which we believe contributed to an increase of the annotation acceptance ratio in our experiments.

5.3 Effectiveness of In-Situ QA Methods

Experiment Setting. We explore two types of in-situ quality assurance methods: annotation time, and redundancy analysis. Annotation time is provided by the annotation platform. As per the acceptance ratio, we only consider the first round of annotation, where the interaction time with reviewers is omitted. For redundancy analysis, we explore two aspects: *intra*-rater Krippendorff's α that measures the consistency of the *same individual* annotator on the same image; and *inter*-rater Krippendorff's α that measures annotation consistency between *different* annotators on the same image. Besides A0 and A2 annotators, we also asked the experienced A1 group to annotate images from the Formal phase. Results in this section are from the first round of annotation.

Annotation Time Analysis. Annotation times for A0 and A2 groups are captured in Table 2 except for the Onboarding phase, where annotators spent time exploring the annotation

[4] https://slack.com

platform and tools. Here, we observe a decreasing trend in the annotation times for A2 annotators, but also see that annotation times have high standard deviation for all annotators regardless of group, indicating variability in each annotator's ability to identify and localize BACs across images.

Continuing to breakdown annotation times based on whether they were accepted or not, a two-sided Welch t-test reveals that a significantly longer annotation time is spent on rejected annotations than accepted ones for both A0 and A2 groups. This observation suggests that when an annotator spends more time annotating an image, it is likely to be difficult to annotate (i.e. the features or patterns are not clearly distinguishable). However, given the wide range of annotation times, it is difficult to rely on this as a quality control method. We may still apply this analysis in the annotation workflow in order to understand if annotators are experiencing difficulties with a certain batch of images, especially when observing longer than average annotation times.

Redundancy Analysis. We show the intra-rater Krippendorff's α results in Figure 5a for A2 annotators, where it is observed that even for the same image, only a small amount of annotations achieve sufficient consistency with high agreement (where Krippendorff's $\alpha > 0.8$). The remaining annotations reveal a large amount of disagreement, which suggests that annotators may not remember their previous annotations on the same image. This is normal for segmentation annotation as it is a *pixel-level* task.

For inter-rater consistency analysis, we compute two different values on images from the Formal phase: consistency between all three groups of annotators shown in Figure 5b, and only between A1 and A2 annotators shown in Figure 5c. In both cases, we observe different distributions of image-level and bounding-box-level Krippendorff's α scores, lending weight to our discussion in Section 3.3 that bounding-box level Krippendorff's α can give rise to more information about annotation inconsistencies for use in review. Furthermore, the discrepancy between bounding-box level and image-level Krippendorff's α scores highlights the difficulty of relying on these as quality assurance methods. Therefore, external mechanisms such as expert review are often necessary for such image segmentation tasks, especially when the visual patterns of ROIs is ambiguous.

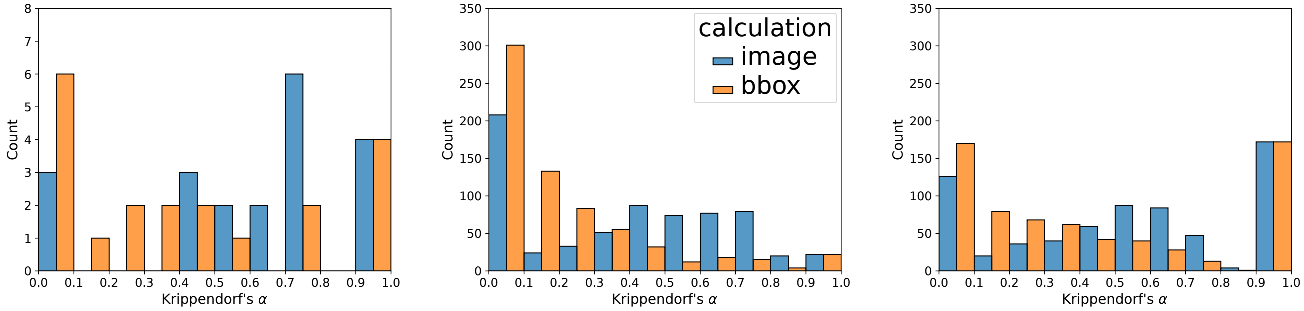
These results also highlight the difficulty due to noise on agreement analysis, in that blindly applying annotation aggregation may result in low-quality outcomes, as the underlying mechanism of many such methods is majority voting.

5.4 Effectiveness of Priority List for Reviewers

Experiment Setting. We now explore how our proposed priority list can help reviewers identify annotations which are likely to have severe quality issues. We use the results from the Formal phase obtained from three annotators with the aforementioned annotation workflow. As there is no established approach for the reviewer's workflow, we consider Random image selection as the baseline. The actual implementation was to assign each image with a random score 10 times with different seeds, and then build a final priority list by averaging the 10 random scores per image then sorting based on these results. We also implement the ranking approaches proposed in Section 4. Note that, one annotation phase can contain multiple annotation iterations, depending on whether an annotation is accepted, and we use the *initial* annotation when creating a priority list. The final output is only used to *evaluate* our priority list in this experiment.

Ground Truth Simulation and Evaluation Methods. Directly letting reviewers provide qualitative evaluation over different ranking lists is difficult. Therefore, to evaluate our results, we adopt approaches often used in evaluating ranking methods by giving a label to each item and then computing common metrics such as Precision (P), Recall (R), and Normalized Discounted Cumulative Gain (N) to understand their effectiveness. Since this approach is still experimental and not implemented, we simulate the ground truth based on whether we believe annotation quality is low enough for reviewers to assess. Our assumption is that annotations are important to review if reviewers have left comments on them, or if the final annotation is very different from the initial annotation. We adopt a graded value for the ground truth because there is a difference in review preference: an annotation tagged as requiring "expert review" has a higher priority when compared to other comments, or no comments. The ground truth label generation process is:

1. Images labelled for "expert review" were scored "3" (36 images).



(a) Intra-rater consistency for A2.

(b) Inter-rater consistency for A0, A1 and A2

(c) Inter-rater consistency for A1 and A2

Figure 5: Inter- and intra-rater annotation consistency analyses with Krippendorff's α.

2. In the remaining images, if reviewers left *any* comments, these were scored “2” (191 images).
3. In the remaining images, those with an image-level Krippendorff’s $\alpha \leq 0.667$ (the default threshold value) were scored a “1” (8 images).
4. All remaining 440 images were scored a “0”.

Analyzing the Priority Lists. We show the evaluation scores at cutoff position $k = 20$ in Table 3, because in the annotation tool, the first 20 images are shown to the reviewers first. More detailed precision and recall at rank positions are shown in Figure 6.

Method	Level	With A0 annotation			Without A0 annotation		
		P@20	R@20	N@20	P@20	R@20	N@20
Random	-	0.30	0.03	0.15	0.30	0.03	0.15
Sorting	Image	0.10	0.01	0.05	**0.75**	**0.07**	0.38
	BBox	**0.60**	**0.05**	**0.31**	0.45	0.04	0.35
Scoring	BBox	0.45	0.04	0.29	0.70	0.06	**0.41**

Table 3: Effectiveness of the various priority list methods.

The worst method is Image-Level Sorting when all three annotations are taken into account, which is worse than the Random method. Since the Image-Level Sorting method heavily depends on image-level Krippendorff’s α scores, its poor performance may be attributed to noisy A0 annotations which reach significantly different conclusions than the bounding-box-level Krippendorff’s α scores. When excluding A0 annotations, Image-Level Sorting becomes much better. The Random method becomes the worst compared to our proposed methods. The sensitive behavior of image-level ranking further confirms our concern with image-level Krippendorff’s α scores, which we believe hides too many details when measuring agreement rates for segmentation tasks such as these. It is worth noting that regardless of whether A0 annotations are included in the ranking process, both bounding-box level approaches are consistently better than the baseline approach, especially when considering NDCG (N) values where graded relevance is used.

Finally, these results suggest that given a fixed review budget, our method can be used to make better use of experts reviewer time. This is highly desirable for complex use-cases where domain experts are expensive and may only afford limited time to spend on review tasks.

6 Conclusions and Future Work

In this paper, we considered the specialized task of crowdsourcing multi-ROI image segmentation annotations in settings where the images for annotation are uncontrolled, highly variable, and where ROI definitions can be ambiguous. Such settings often require crowdsourced annotations to be reviewed by domain experts to ensure high quality annotations are ultimately produced. There are many such real-world settings which fall into this category, including high-stakes medical imaging domains, or environmental remote sensing applications where precise image segmentation annotations are required to build accurate image segmentation models.

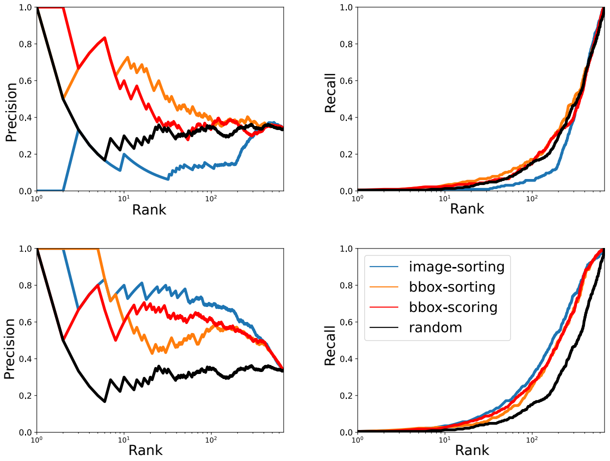


Figure 6: P@k (the 1st column) and R@k (the 2nd column) ($k = \{1, 2, \ldots, 675\}$) of the priority list with three annotations (the 1st row) and only A1, A2 annotations (the 2nd row). The x-axis is log-scaled.

We have shown that basic annotation workflows with minimal quality assurance can result in low quality annotations. The interpretation of complex images with ambiguous ROIs is not straightforward, and relying on crowdsourced annotators to precisely localize ambiguous ROIs can result in high variability, hindering techniques such as annotation fusion. We also observed that the same annotator may not reliably reproduce the same annotations for the same image, and adding more annotators is simply likely to introduce more disagreement. Our proposed annotation workflow extends the basic annotation workflow by introducing phases which allow task owners to iteratively improve the task settings and annotator understanding through feedback to avoid mistakes. We have demonstrated through our experiments that this new workflow provides us with a better acceptance ratio for finalized image annotations, as well as higher quality annotations overall.

Our experimental results also suggest that our proposed methods of prioritizing annotations for review are robust to noise or error, and the priority list generated with our methods ranks “interesting” annotations highly to ensure that a suitable trade-off between reviewer effort and annotation quality can be met. We expect that our methods for prioritizing images for review can be applied to other image segmentation annotation tasks outside the medical imaging domain.

In the future, we plan to extend our work in the following ways: (i) incorporating automated methods of image selection in the annotation workflow using active learning. Such methods aim to select new images on the basis that their annotation will be informative to a model in re-training to improve its quality or performance; and (ii) reducing the effort of annotators by incorporating trained deep segmentation models to generate pre-annotations to speed up the annotator training and labelling processes.

References

Amgad, M.; Elfandy, H.; Hussein, H.; Atteya, L. A.; Elsebaie, M. A.; Abo Elnasr, L. S.; Sakr, R. A.; Salem, H. S.; Ismail, A. F.; Saad, A. M.; et al. 2019. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proc. ACM CHI*, 2334–2346.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *Comput. Surveys*, 51(1): 1–40.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 248–255.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proc. ACM CSCW*, 1013–1022.

Freeman, B.; Hammel, N.; Phene, S.; Huang, A.; Ackermann, R.; Kanzheleva, O.; Hutson, M.; Taggart, C.; Duong, Q.; and Sayres, R. 2021. Iterative quality control strategies for expert medical image labeling. In *Proc. AAAI HCOMP*, volume 9, 60–71.

Gurari, D.; Theriault, D.; Sameki, M.; Isenberg, B.; Pham, T. A.; Purwada, A.; Solski, P.; Walker, M.; Zhang, C.; Wong, J. Y.; and Betke, M. 2015. How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-experts, and Algorithms. In *Proc. IEEE WACV*, 1169–1176.

Highnam, R.; Brady, S. M.; Yaffe, M. J.; Karssemeijer, N.; and Harvey, J. 2010. Robust Breast Composition Measurement - Volpara™. In *Digital Mammography*, 342–349. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-13666-5.

Iribarren, C.; Chandra, M.; Lee, C.; Sanchez, G.; Sam, D. L.; Azamian, F. F.; Cho, H.; Ding, H.; Wong, N. D.; and Molloi, S. 2022. Breast Arterial Calcification: a Novel Cardiovascular Risk Enhancer Among Postmenopausal Women. *Circulation: Cardiovascular Imaging*, 15(3).

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 12341–12351.

Jones, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jungo, A.; Meier, R.; Ermis, E.; Blatti-Moreno, M.; Herrmann, E.; Wiest, R.; and Reyes, M. 2018. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Proc. MICCAI*, 682–690.

Kazai, G.; and Zitouni, I. 2016. Quality management in crowdsourcing using gold judges behavior. In *Proc. WSDM*, 267–276.

Kovashka, A.; Russakovsky, O.; Li, F.; Grauman, K.; et al. 2016. Crowdsourcing in computer vision. *Found. Trends Comput. Graph. Vis.*, 10(3): 177–243.

Krause, J.; Gulshan, V.; Rahimy, E.; Karth, P.; Widner, K.; Corrado, G. S.; Peng, L.; and Webster, D. R. 2018. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8): 1264–1272.

Krippendorff, K. 2011. Computing Krippendorff's Alpha-Reliability. https://repository.upenn.edu/asc_papers/43. Accessed: 2022-12-20.

Kutlu, M.; McDonnell, T.; Barkallah, Y.; Elsayed, T.; and Lease, M. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement? In *Proc. SIGIR*, 805–814. New York, NY, USA: ACM.

Lampert, T. A.; Stumpf, A.; and Gançarski, P. 2016. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Trans. on Image Processing*, 25(6): 2557–2572.

Lee, S. C.; Phillips, M.; Bellinge, J.; Stone, J.; Wylie, E.; and Schultz, C. 2020. Is breast arterial calcification associated with coronary artery disease? – A systematic review and meta-analysis. *PLOS ONE*, 15(7): 1–19.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*, 740–755.

Ørting, S. N.; Doyle, A.; van Hilten, A.; Hirth, M.; Inel, O.; Madan, C. R.; Mavridis, P.; Spiers, H.; and Cheplygina, V. 2020. A Survey of Crowdsourcing in Medical Image Analysis. *Human Computation*, 7: 1–26.

Pei, W.; Yang, Z.; Chen, M.; and Yue, C. 2021. Quality Control in Crowdsourcing based on Fine-Grained Behavioral Features. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2): 1–28.

Ribeiro, V.; Avila, S.; and Valle, E. 2019. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proc. ACM CHI*, 1–15.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. SIGKDD*, 614–622.

Singhal, A.; Buckley, C.; and Mitra, M. 1996. Pivoted document length normalization. In *Proc. SIGIR*, 21–29.

Waggoner, B.; and Chen, Y. 2014. Output agreement mechanisms and common knowledge. In *Proc. AAAI HCOMP*.

Wang, J.; Ding, H.; Bidgoli, F. A.; Zhou, B.; Iribarren, C.; Molloi, S.; and Baldi, P. 2017. Detecting cardiovascular disease from mammograms with deep learning. *IEEE Trans Med Imaging*, 36(5): 1172–1181.

Wang, K.; Hill, M.; Knowles-Barley, S.; Tikhonov, A.; Litchfield, L.; and Bare, J. C. 2022. Improving Segmentation of Breast Arterial Calcifications from Digital Mammography: Good Annotation Is All You Need. In *Proc. ACCV workshops*, 130–146.

Warfield, S. K.; Zou, K. H.; and Wells, W. M. 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, 23(7): 903–921.